MEASUREMENTS THAT SIMULATE HUMAN HEARING – PSYCHOACOUSTICS IN ITU-T P.863

ROHDE&SCHWARZ

Make ideas real

MEASUREMENTS THAT SIMULATE HUMAN HEARING – PSYCHOACOUSTICS IN ITU-T P.86

Even though telephony now utilizes only a small part of network resources, it remains a core function of telecommunications networks if measured in terms of the actual duration of usage – and speech quality is an important criterion for the acceptance of a service.

Automatic measurement procedures can be used to assess just how well an end-to-end connection is performing.

These procedures involve meticulously characterizing human hearing based on psychoacoustic models.

This is shown here using the state-of-the-art ITU-T P.863 standard as an example.

Almost everyone is familiar with technical parameters such as signal-to-noise ratio, total harmonic distortion and frequency response, which are used to provide insight into the sound quality of high fidelity audio equipment. Until the 1990s, purely physical parameters of this type were used for technically evaluating of the quality of telephone connections. When using analog transmission or simple PCM methods, these types of parameters were also adequate to allow a rough estimate of the transmission quality. Even after the first coding methods came into use – still with the goal of delivering the most exact reproduction possible of the waveform (DPCM, ADPCM) – measurements were still largely restricted to capturing the differences in amplitude between the transmitted signal and the original signal. The channel was simplified by modeling it as a time-invariant, linear system for speech transmission, and any deviation from this assumption was treated as additive distortion.

These assumptions have come under pressure with the introduction of new speech coding methods. Code-excited linear prediction (CELP) as well as frequency-domain coding methods were optimized for high acceptance of the coded audio or speech signal – and no longer necessarily for nearly exact reproduction of the signal as a waveform. Consequently, amplitude differences between the input and output signals could not be generally regarded as perceptible qualitative distortions of the speech signal.

It was at this time that the first psychoacoustic motivated speech quality measurement algorithms were created; the current ITU-T standard P.863 POLQA can be considered the most successful and precise representative of these algorithms.

The goal is to produce technically derived quality assessments of transmitted speech signals that are comparable to estimations obtained in auditory tests using test persons. The speech quality measurement procedure evaluates the quality of a short speech on a scale just like a large group of listeners would. Put simply, the quality is rated using a five-level, one-dimensional mean opinion score (MOS) and the opinions of all test persons are averaged. Now the speech signal has to be technically analyzed in order to calculate this single value for speech quality. Psychoacoustic motivated methods model the auditory situation, human auditory physiology and speech perception in order to attain as precise a result as possible.

Test scenario for speech quality measurements

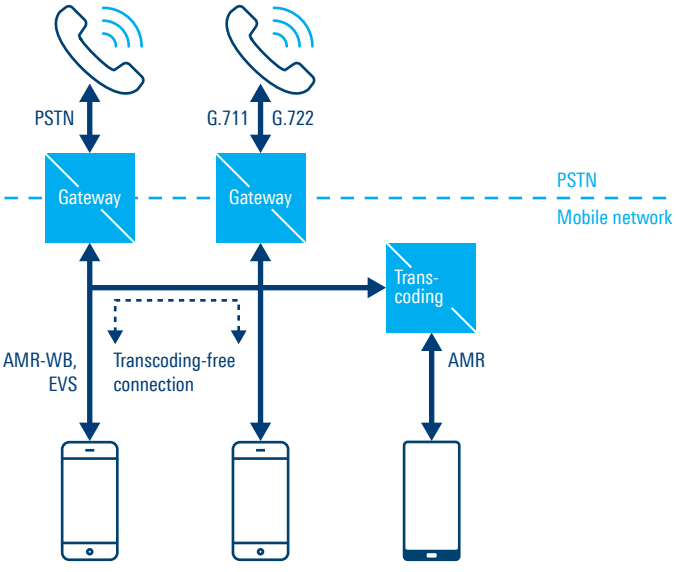


Fig. 1: Typical end-to-end test connections for speech quality measurements.

The basic structure of ITU-T P.863

ITU-T P.863 approach is called full-reference, which means that the quality prediction is based on the comparison between an undistorted reference (original) signal and the received signal. To evaluate mobile communications channels, a reference speech signal from a preconfigured remote station is received, recorded and compared with a local copy of the reference signal; this is a typical end-to-end measurement. The most common use cases involve test calls between two mobile phones or from a mobile phone to a landline connection (Fig. 1). The measuring systems are designed primarily for mobile use in order to measure the quality of speech connections in real networks while on the move. Of course, P.863 can also be used for evaluation of pure landline or VoIP connections as well as in the lab.

In simplified terms, ITU-T P.863 – like its predecessor ITU-T P.862 – has three aspects:

- Preprocessing and synchronization of the reference and test signals
- Modeling of auditory physiology
- Modeling of speech perception and temporal integration

All of the analyses as well as the calculation of the quality value are based exclusively on the speech signal itself. ITU-T P.863 does not require any additional information or even IP data. This allows very wide-ranging applications since no knowledge is needed about the transmission system at measurement runtime; the transmission channel is treated as a black box (Fig. 2).

Transmission path and modeling in ITU-T P.863

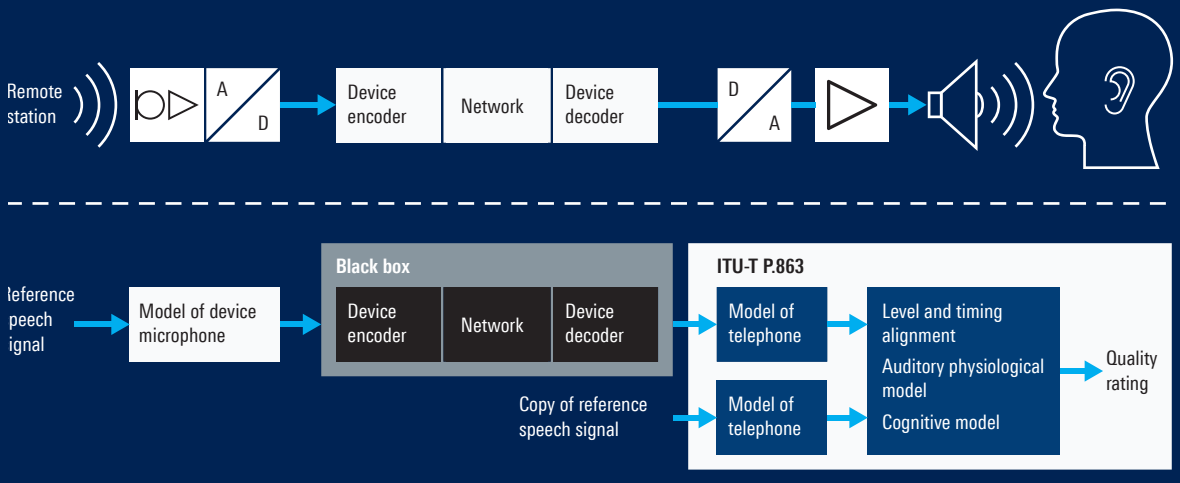


Fig. 2: The transmission channel is treated as a black box.

Preprocessing and synchronization of the reference and test signals

Preprocessing of the signals includes some basic technical aspects such as correcting deviating clock frequencies during transmission, adjusting the signal level and prefiltering the signals as needed in order to model the auditory situation, i.e. with the frequency response of a headset or – in narrowband mode – with the frequency response of a typical telephone receiver. However, the most important and challenging part of preprocessing involves precise synchronization of the reference and test signals to allow subsequent analysis in brief time windows.

The measurement procedure – which will be discussed later – is based on a comparison of the internal frequency-time representations of the reference and transmitted signals. It must be possible to congruently compare the two representations on the time axis, i.e. a certain segment of the signal under evaluation $y(t + \tau)$ must be comparable with the corresponding reference segment $x(t + \tau)$. In the past, it was reasonable to assume a constant delay in the transmission system, i.e. a constant time offset between the two signals $x(t + \tau + c) \sim y(t + \tau)$. In that case, a simple correction using the constant c sufficed to synchronize the representations on the time axis. For largely linear systems, c could be easily determined and with sufficient precision by analyzing the cross-correlation between the input and output signals.

Today's transmission systems no longer meet these prerequisites because they are time-variant and nonlinear. The strong time variance is obvious in popular Internet voice communications services with their perceptible fluctuation in speech rate. At times the speech tempo appears compressed, and at other times elongated. Nevertheless, to be able to precisely compare the short-term spectra of the reference signal and the signal under evaluation, all of the equivalent segments of the two signals must be aligned.

ITU-T P.863 does this with a multi-stage iterative procedure. First, the signals are prealigned in the time domain. The initial delay is eliminated, the signal is divided into long segments such as sentences or word groups, and these are roughly synchronized. Based on this rough alignment, the individual sections are subdivided more finely and then aligned exactly.

It is assumed that the two signals are correlated at least over short time intervals, if not in the time domain then definitely in the frequency domain. Therefore, both spectral similarities and cross-correlations are used as the synchronization criteria. The signal under evaluation is iteratively broken into smaller and smaller segments and matched with the corresponding parts in the reference signal until the synchronization criteria are fulfilled. At the end of the synchronization procedure, each segment in the signal under evaluation has been aligned with the corresponding segment in the reference signal.

When all speech segments have been aligned, simple desynchronization is checked by repeatedly and uniformly shifting the signals – such as can occur when there are deviations in the clock frequencies between the transmitting and receiving ends. This type of desynchronization also leads to drift in the frequency domain. If necessary, this is corrected prior to further processing by resampling the signal – a complicated process. The alignment procedure is then repeated to check if the resampling was successful.

In most cases, all of the signal segments can be synchronized. But there can also be highly distorted, missing or inserted (artificial) signal segments. These cannot be reliably aligned with a segment in the reference signal. Such segments are interpolated between accurately assigned segments based on plausibility criteria or they are added to these segments. At the end of the synchronization process, it is ensured that even the tiniest received signal segment can be assigned to a corresponding segment in the reference signal. The signals are not reproduced as a synchronized time domain signal. Instead, the alignment is represented virtually using a correspondence table that contains the start and end points of the matching signal segments (Fig. 3).

Based on this alignment, the two signals are synchronously mapped in overlapping windows (FFT) and psychoacoustically transformed to produce a hearing-oriented, time-frequency representation.

Auditory physiology and speech perception

First, the basics of auditory physiology will be outlined, i.e. the transformation of a sound event into an internal stimulus. This represents the main component of the speech-processing model in ITU-T P.863.

Certain hearing phenomena are generally known. One well-known phenomenon is the absolute threshold of hearing under which sounds cannot be perceived. It is frequency-dependent – as is human perception of sound intensity. An example is the well-known A-weighting curve, which is a frequency weighting for a certain sound intensity. The perception of volume and tone pitch, which is roughly logarithmic, is taken into account by using a decibel scale for the intensity and listing by octave or third-octave bands in the frequency domain. However, the underlying physiology of hearing is much more complex. In simplified terms, the goal of psychoacoustic motivated algorithms is to provide a signal as a starting point for quality estimation – like the signal delivered by the auditory nerve to the speech-processing center in the human brain.

The fundamental idea of the ITU-T P.863 measurement algorithm is not to calculate the difference between the original signal (reference signal) and the transmitted, distorted signal at the level of measurable amplitude values. Instead, it is based on differences in the "internal representation" of the signals, i.e. the actual signal available to the human brain after the speech signal has undergone auditory physiological processing. Simply put, this means masking out signal components that are inaudible.

How is the transformation of the sound signal into an internal stimulus modeled? ITU-T P.863 begins with a short-term spectral analysis. The speech signal is subdivided into overlapping windows with lengths of 30 ms to 40 ms and converted into a spectral display using FFT, whereby the sound pressure level is normalized to what is known as the ear reference point (ear level). The result of the FFT is an equidistant spectral representation of the sound energy in the frequency domain. The spectral resolution of sound events in the inner ear is not constant across the audible frequency range. It becomes less clear as the frequency increases. In physiological terms, this is due to what is known as the frequency-to-place transformation on the basilar membrane in the inner ear. The basilar membrane, which is coiled up in the cochlea of the inner ear and immersed in liquid, is the base for the sensory hair cells. The membrane is stimulated via the ossicles: the stapes, malleus and incus. Depending on the stimulus frequency, vibration maxima are formed closer to the point of stimulus or further inside closer to the rear support of the membrane. In this way, the stimulus frequency vibrates at a certain place on the basilar membrane. Since the membrane is narrower towards the inside, it becomes stiffer and adjacent frequencies lead to more closely spaced vibration maxima than at the front end of the membrane. Since the density of the hair cells is approximately constant, the spectral resolution of human hearing decreases at higher frequencies.

Comparison of reference signal and test signal

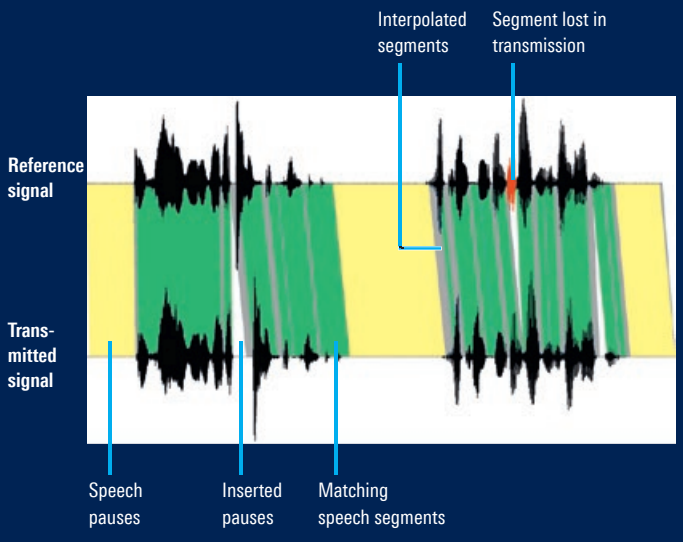


Fig. 3: Example of timing alignment of speech segments in the reference and test signals in case of time-variant transmission.

This transformation of an air pressure fluctuation by the eardrum and the ossicles into a pressure variation in the cochlear fluid can be interpreted as an acoustic impedance conversion. This is the only way that an acoustic air vibration with a wavelength between 15 m and about 1.5 cm in the range of human hearing can be mapped on a liquid-supported membrane with a length of approx. 7 cm (Fig. 4).

This basic understanding of the transformation of an acoustic sound event into a stimulus of the auditory nerve explains many phenomena related to hearing. The hair cells require a minimum deflection in order to respond. This, in conjunction with a noise floor, forms the resting threshold. The nonequidistant frequency-to-place transformation on the narrowing basilar membrane is responsible for the approximately logarithmic relationship between frequency and perceived tone pitch. The Bark scale, which has been around for a long time, does a good job of modeling this situation.

Considering the described frequency-to-place transformation and the vibration of the basilar membrane, it becomes clear that a single frequency (a sinusoidal oscillation) does not cause a deflection of the membrane at just a single position; instead, a whole region of the membrane is vibrated, whereby only its maximum corresponds to the excitation frequency. This means that adjacent hair cells are also vibrated and stimulated.

This stimulus of an entire region leads to the well-known phenomenon of spectral masking. Weak stimuli in the immediate vicinity of a stronger stimulus do not cause (additional) perceptible excitation of the affected sensory cells. The weak stimulus is masked by the stronger adjacent stimulus and cannot be perceived (or only to a limited extent). One of the achievements of human hearing is that despite the excitation of a whole region on the basilar membrane, a sinusoidal signal can be perceived as such – and not as narrowband noise. However, this comes at the cost of reduced sensitivity in the area of this excitation maximum.

Besides spectral masking, there is another effect known as temporal masking. Following an intensive sound event, the sensitivity of the hair cells in the excited region is reduced for a brief time interval. Weaker stimuli will not be perceived there (or only weakly perceived) for some milliseconds (see box).

Masking phenomena in coding

The phenomenon of spectral and temporal masking was described by Zwicker in the 1950s. This model was not widely used until the development of the MP3 audio coding method, and it was instantly a resounding success. MP3 reduces the signal components that human hearing automatically masks. The coding distortions are hidden below the masking threshold and are therefore imperceptible or barely perceptible. MP3 and later methods such as AAC and WMA exploit the redundancy of human hearing to reduce the amount of information transmitted with the least possible impairment. There is also a distinction between audio and speech coding methods. Speech coding methods exploit the redundancies of speech produced by the human vocal tract. The resulting speech signal is highly correlated and can be characterized reasonably well using simple predictive models. State-of-the-art coding methods model the principle of how the vocal chords and vocal tract produce speech, transmit the model parameters and synthesize the speech signal on the receiving end. This pure vocoder method became commercially viable when, in addition to the original speech signal, the error signal of the synthesized signal was transmitted and used on the receiving end to correct the generated, artificial speech signal. Further development of this coding approach primarily focuses on increasing the efficiency of transmitting as precise an error signal as possible.

Anatomy of the human ear

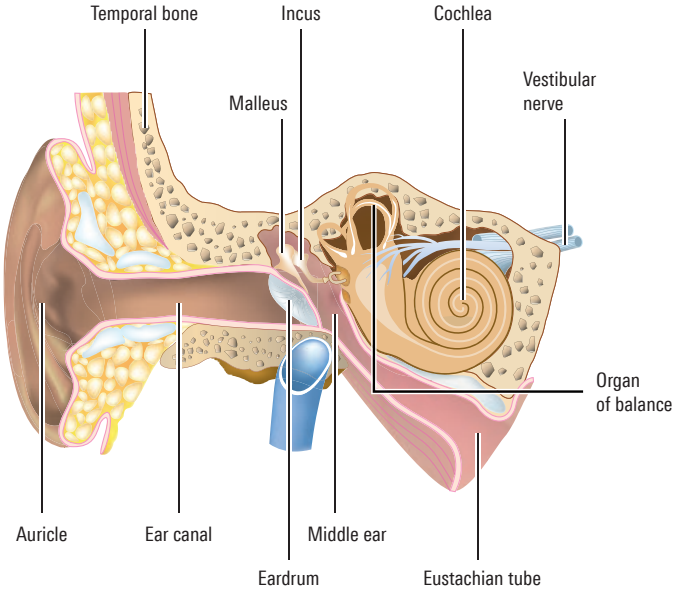


Fig. 4: The anatomy of the human ear. The frequency range is logarithmized in the cochlea.

Modeling of these spectral and temporal masking effects is the next and most important step in the transformation of the original signal spectrum into a perception-oriented internal representation. The transformed signal can be understood as a special form of a spectrogram in which the frequency-time representation is characterized only by sound intensities that are actually perceptible (sonogram).

The measurement procedure now sees this frequency-time representation as an internal representation of the signal. By comparing the two "spectrograms" of the reference signal and the transmitted signal, a difference representation is obtained that only shows the differences that are actually perceptible. Signal components that are not perceived due to auditory physiology no longer play a role here.

Speech perception and quality modeling

At first glance, it may seem that simply accumulating the perceptible difference vs. frequency and time would lead to a perception-oriented disturbance rating and a valid quality assessment. However, this is not the case.

On the one hand, this "difference" characterizes only the perceptible difference under the remote assumption that both signals can be directly compared and evaluated. However, even a delayed presentation leads to an increased error tolerance since even though acoustic differences lead to different physiological stimuli, they are not remembered as such. The error tolerance for a telephone connection (the case in this model) is higher because the listener can only compare the spoken words with their memory of the speaker as well as their expectations based on their experience of how human speech naturally sounds. Even significant measurable deviations are considered irrelevant, leading to a further masking of errors that is more difficult to comprehend.

As has already been discussed, the goal of the speech quality measurement procedure is to realistically characterize the perception of a speech signal in terms of its naturalness and the influence of disturbances. However, the range of what is acceptable as natural speech is quite large. People are accustomed to perceiving many different voices as natural. People's tolerance to voice is much higher than for a musical instrument. If an instrument is not perfectly tuned, the listener notices immediately. This tolerance in speech perception means that not all theoretically perceptible differences to the reference signal will affect the quality rating equally. Slight shifts in the fundamental frequency and formants tend to be better tolerated than additive distortion caused by noise or interference pulses. There is also a higher tolerance to ripple in frequency responses or even to band restrictions as well as to slow, steady level changes or even to changes in the temporal structure of the signal, i.e. slight elongation or compression on the time axis.

Postprocessing of the calculated, perceptible signal differences is necessary based on the analytical power of the brain. Basically, the auditory physiology only modeled the organic hardware; now the hearing software has to be added to the model. This can be done with a postprocessing cognitive model approach in addition to the described auditory physiological model. In contrast to the precise algorithmic model of the inner ear, the cognitive model is implemented rather coarsely in today's models and is limited to specific weightings of individual error patterns, which in turn are assigned to individual, specific causes (e.g. frequency response, additive noise, etc.).

The model in ITU-T P.863 divides the perceptible signal differences into four categories (indicators), which are analyzed independently and then later weighted and included in the overall assessment:

- (Additive) noise
- Frequency response / timbre
- Reverberation
- Distortions

Simplified structure of ITU-T P.863

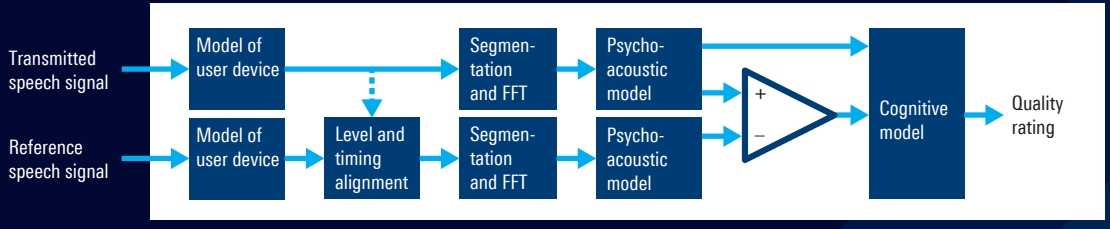


Fig. 5: The structure of ITU-T P.863 POLQA.

Distortions represent the most important branch of analysis. The differences in the sonograms for the reference and test signals over all speech components are used as the basis for calculation. Before these differences are calculated, however, the reference signal representation is aligned with the test signal; level fluctuations and deviations in the timbre are partially corrected in this process. Even though they barely impact the perceived quality, they would dominate the calculated differences. When accumulating the absolute differences, a distinction is made between missing and added elements. Missing signal intensities have a much lower weighting. The accumulated differences over the individual frequency ranges and speech components represent the baseline for the quality rating.

Since the influence of the timbre (frequency response) of the speech signal on the disturbance value was reduced prior to calculation, this indicator is taken into account in a separate evaluation. Reverberation in the speech signal is also quantified separately. Although a certain amount of reverberation in the speech signal hardly leads to lower perceived quality, it does lead to a greater difference in the spectral display and does not correspond to what is perceived. Both indicators – timbre and reverberation – are used to correct the calculated base value for the quality.

This calculated quality base value does a relatively good job of characterizing speech perception, but additive noise is mainly perceptible during speech pauses when there is no speech signal to mask it. This is why background noise is measured separately, weighted and also used to adjust the quality value. The quality model is finalized by transforming the calculated quality value to the MOS scale of 1 to 5. This transformation is based on listening tests with test persons; their MOS values need to be reproduced as precisely as possible by the measured value of the model (see box).

Model development and test data

ITU-T P.863 is the result of a multi-year competition organized by the ITU for a new standard of speech quality evaluation in telecommunications networks. In particular, the advance of IP based transmission and new codecs as well as the extension of the transmitted speech spectrum to HD voice and beyond have necessitated a new measurement method.

Following a complex selection process, the models created by OPTICOM, TNO and Rohde&Schwarz SwissQual AG were chosen as the winners of the competition. Combining the advantages of the three successful candidates into a single model to be standardized led to ITU-T P.863 POLQA. Due to synergy effects, the new model considerably exceeded the accuracy of the individual approaches and was officially approved as a standard in 2010. More than 45000 spoken and transmitted sentences in ten different languages that were evaluated by listeners were used for the selection and verification of ITU-T P.863.

Today, ITU-T P.863 is the reference method for speech quality measurements. It has been installed on many thousands of measuring instruments worldwide for quality monitoring purposes. The standard is kept up-to-date through regular model maintenance. Published in March 2018, version 3.0 was specially verified for the new EVS coding method. It now also supports the entire audio spectrum audible to humans.

ITU-T P.863 in Rohde&Schwarz products

Rohde&Schwarz SwissQual AG – a partner in the POLQA coalition and a co-owner of the rights – has belonged to Rohde&Schwarz since 2012. The long-standing experience of this company in the area of speech and video analysis and quality modeling has been integrated into the Rohde&Schwarz product portfolio. Core products are QualiPoc and Benchmarker – both are smartphone based measuring systems that provide all relevant information at the RF and IP levels and analyze the media signal (audio or video) in real time. Call setup and release, feeding and recording of speech signals, and saving of result structures is automatic. The systems are designed for extensive series measurements in real networks.

After the ITU-T P.863 based algorithm was approved as a standard, it was immediately implemented under Android. Thanks to its own model development, QualiPoc had a head start and was the first commercially available system to run P.863 on a smartphone in real time (Fig. 7).

P.863 POLQA is the core element of an enhanced audio analysis. In addition to the MOS value, other technical parameters such as the speech and noise level, but also important parameters in today's telecommunications channels such as "missed voice" (the percent of lost, unreceived speech) are calculated and output.

Fig. 6: Vehicle based test campaigns can be used to prepare comprehensive mobile coverage and quality maps. Speech quality measurements are part of the procedure.

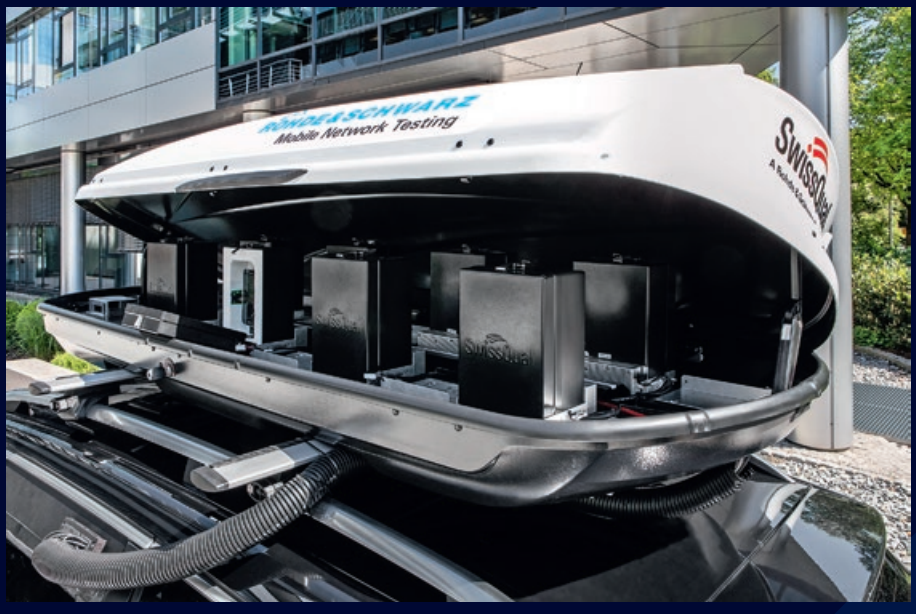
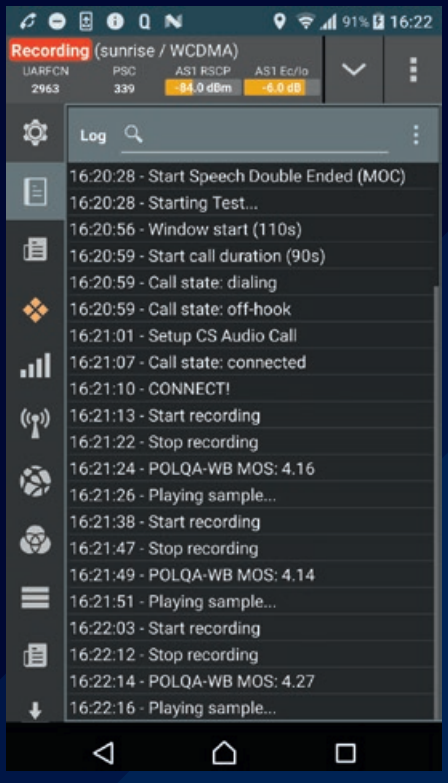


Fig. 7: Real-time evaluation of a speech signal in line with ITU-T P.863 using a QualiPoc measuring system.



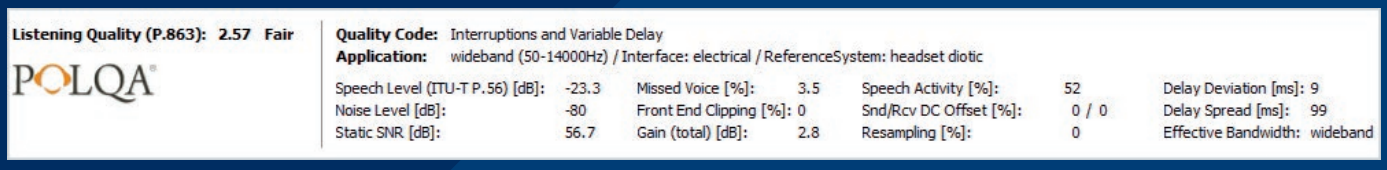
Many of the parameters can be visualized over the signal duration. This allows errors in the speech signal to be precisely localized and synchronized with the lower transmission layers, which enables detailed error analysis (Fig. 8).

Information about disturbances that influence the quality value allows automatic analysis of the root causes behind a decrease in speech quality. This type of root cause analysis can distinguish among almost 30 different primary sources of problems, including background noise, heavy coding distortion, lost packets, interruptions and more. This auxiliary information is very helpful in pinpointing the problem on the transmission path.

Today, over three quarters of all measuring systems sold include optional audio analysis – a clear indication of the degree of acceptance and success of automated speech quality measurements. However, usage of ITU-T P.863 is not limited to evaluation of classic mobile phone calls. The focus on evaluating the speech signal also allows assessment of OTT services such as WhatsApp Call and Skype. In recent years, it has become increasingly important for network operators to compare these services with their own products in the field, thereby contributing to the high demand for suitable measuring systems.

Dr. Jens Berger

Fig. 8: Example of speech quality analysis with P.863. The MOS value for the speech signal is 2.57 (overall rating: "Fair"). Interruptions and variable delay are given as the reason for the rather low score. This score is quantified by the 3.5 % of lost speech signals (missed voice) and a delay spread of 99 ms.



Service at Rohde&Schwarz
You're in great hands

- Worldwide
- Local and personalized
- Customized and flexible
- Uncompromising quality
- Long-term dependability

Rohde&Schwarz

The Rohde&Schwarz technology group is among the trailblazers when it comes to paving the way for a safer and connected world with its leading solutions in test&measurement, technology systems and networks&cybersecurity. Founded more than 85 years ago, the group is a reliable partner for industry and government customers around the globe. The independent company is headquartered in Munich, Germany and has an extensive sales and service network with locations in more than 70 countries.

www.rohde-schwarz.com

Sustainable product design

- Environmental compatibility and eco-footprint
- Energy efficiency and low emissions
- Longevity and optimized total cost of ownership

Certified Quality Management
ISO 9001

Certified Environmental Management
ISO 14001

Rohde&Schwarz training

www.training.rohde-schwarz.com

Rohde&Schwarz customer support

www.rohde-schwarz.com/support

PD 5216.0721.32 | Version 02.00 | January 2023
Measurements that simulate human hearing – psychoacoustics in ITU-T P.863
Data without tolerance limits is not binding | Subject to change

5216.0721.32 02.00 PDP 1 en